A picture is worth a thousand words: Using OpenClipArt library to enrich IndoWordNet

Diptesh Kanojia, Shehzaad Dhuliawala, and Pushpak Bhattacharyya
Centre for Indian Language Technology,
Computer Science and Engineering Department,
IIT Bombay,
Mumbai, India
{diptesh,shehzaadzd,pb}@cse.iitb.ac.in

Abstract

WordNet has proved to be immensely useful for Word Sense Disambiguation, and thence Machine translation, Information Retrieval and Question Answering. It can also be used as a dictionary for educational purposes. The semantic nature of concepts in a WordNet motivates one to try to express this meaning in a more visual way. In this paper, we describe our work of enriching IndoWordNet with image acquisitions from the OpenClipArt library. We describe an approach used to enrich WordNets for eighteen Indian languages.

Our contribution is three fold: **(1)** We develop a system, which, given a synset in English, finds an appropriate image for the synset. The system uses the OpenclipArt library (OCAL) to retrieve images and ranks them. **(2)** After retrieving the images, we map the results along with the linkages between Princeton WordNet and Hindi WordNet, to link several synsets to corresponding images. We choose and sort top three images based on our ranking heuristic per synset. **(3)** We develop a tool that allows a lexicographer to manually evaluate these images. The top images are shown to a lexicographer by the evaluation tool for the task of choosing the best image representation. The lexicographer also selects the number of relevant images. Using our system, we obtain an Average Precision (P @ 3) score of 0.30.

1 Introduction

Our goal is to enrich the semantic lexicon of various Indian languages by mapping it with images from the OpenClipArt library(Phillips, 2005). India is currently experiencing a major enhancement in the digital education sector with its vision of the 'Digital India' program[1]. In this paper, we introduce an approach to enrich the IndoWordNet[2](Bhattacharyya, 2010), with images, which can help students and language enthusiasts alike. We envision the use of WordNets in the education sector to promote language research among young students, and provide them with a multilingual resource which eases their study of languages. WordNets have proven to be a rich lexical resource for many NLP sub tasks such as Machine Translation (MT) and Cross Lingual Information retrieval.

India has 22 official languages, written in more than 8 scripts. When a user reads a concept in a language that is not known to them, and moreover in an unknown script, an image can provide helpful insight into the concept. Language learners in a multilingual country like this often face difficulty mainly due to: **(a)** Not being able to find a mapping of the concept in the language being studied and their native language and **(b)** Not being able to decipher the script in the language being learnt. In such cases a pictorial representation of a concept will be very useful.

Finally, systems for Automatic image captioning and Real time video summarization can leverge the power of image enriched WordNets.

[1]http://www.digitalindia.gov.in/
[2]http://www.cfilt.iitb.ac.in/indowordnet

1.1 WordNets and IndoWordNet

WordNets are lexical structures composed of synsets and semantic relations(Fellbaum, 1998). Such a lexical knowledge base is at the heart of an intelligent information processing system for Natural Language Processing and Understanding. IndoWordNet is one such rich online lexical database containing more than twenty thousand parallel synsets for eighteen languages, including English. It uses Hindi WordNet as a pivot to link all these languages. The first WordNet was built in English at Priceton University[3]. Then, followed the WordNets for European Languages[4](Vossen, 1998), and then IndoWordNet. IndoWordNet has approximately 25000 synsets linked to Princeton WordNet. We use these linkages to mine English words from the Princeton WordNet which form the basis of our query for the OpenClipArt API. We download the images via their URLs, and store them locally, to map them to Hindi WordNet[5](Dipak Narayan and Bhattacharyya, 2002) synset IDs later.

The paper is organized as follows. In section 2, we describe our related work. In section 3 and 4, we describe our architecture, and the retrieval procedure along with the scoring algorithm. We describe the results obtained in Section 5. We describe the evaluation tool and qualitative analysis in sections 6 and 7, respectively. We conclude in section 8.

2 Related Work

Bond et al. (2009) used OCAL to enhance the Japanese WordNet, and were able to mine 874 links for 541 synsets. On the basis of manual scoring they found 62 illustrations which were best suited for the sense, 642 illustrations to be a good representation, and 170 suitable, but imperfect illustrations. We extend their work for IndoWordNet, and use OCAL to mine the illustrations. Imagenet(Deng et al., 2009) is a similar project for Princeton WordNet which provides images/URLs for a concept. It contains 21841 synsets indexed with 14,197,122 images. We present a much simpler methodology of collecting images from the web, and then using the synset words to find overlaps with image tags, and then map them.

[3]http://www.wordnet.princeton.edu
[4]http://www.illc.uva.nl/EuroWordNet/
[5]http://www.cfilt.iitb.ac.in/wordnet/webhwn/

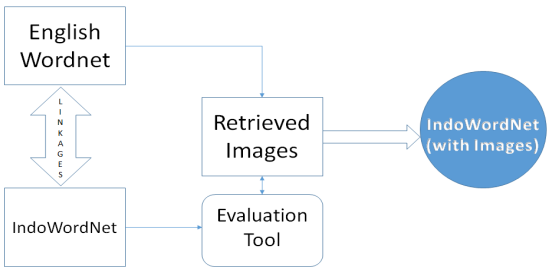

Figure 1: Our Architecture

3 Our Architecture

The following section gives the detailed architecture of our system. A diagrammatic representation is shown in figure 1. Also, we discuss the structure of the IndoWordNet and talk about how we link it to the retrieved set of images.

3.1 Dataset

A linked Hindi - English synset mapping is required to mine the image-synset mapping for Hindi. OpenClipArt contains URL tags in English, and thus a linked Hindi - English synset data structure was required. For our work, we use the following data sets:

3.1.1 Hindi Database

The latest version of Hindi WordNet is available for download at: `http://www.cfilt.iitb.ac.in/wordnet/webhwn/downloaderInfo.php`, which provides an offline interface along with the database, in text format.

3.1.2 English Database

The latest version of Princeton WordNet is available for download at: `https://wordnet.princeton.edu/wordnet/download/`. It provides both the latest database, and standalone installers for WordNet

3.1.3 Hindi-English Linkage database

WordNets have been built for around 100 different languages. Efforts towards mapping synsets across WordNets have been going on for a while in various parts of the world. IndoWordNet contains 28,446 synsets linked to the Princeton WordNet, out of which 21,876 are Nouns. Those concepts in Hindi for which there are no direct linkages in the English WordNet, it was decided to link them to a

hypernymy synset in English. The idea was that instead of having no linkage at all there would be at least a super-ordinate concept and lexical item/items with which the Hindi concept could be linked to provide *weak* translation candidates which could be exploited for various NLP tasks. IndoWordNet has 11,582 direct linkages, and 8184 hypernymy linkages. We use only 11,582 directly linked noun concepts to mine OCAL.

4 Retrieval procedure and scoring

We use the OpenClipArt API[6] to retrieve a set of results using the head word from a synset as the query, since OpenClipArt is a free to use resource, unlike Google Search results which might retrieve copyright data. The API provides a JSON output which can be easily parsed using any programming language. We use JAVA for this purpose. The result for each image provides the following data:

- The title of the 'image'
- The tags for the 'image'
- The URL of the 'image'

To rank the results, we calculate a score based on overlaps between the synsets and image meta-data. The score is derived as a weighted overlap between the words in the *Title* and *Tags* of the result image with the words of the synset. Words from each part are given a different weight owing to how useful the feature is in describing the image. For example, words from the Title are given a higher weight as compared to words from the image Tags. The algorithm increases the score if an overlap occurs and decrements the score otherwise. The magnitude of this increase and decrease depends on the weights of the words being compared. Our system allows for all these weights to be tweaked.

After the result images are scored, they are sorted based on this score. Only the top three scoring images are downloaded. These downloaded images are then evaluated by lexicographers.

5 Results

Using the methodology described above, we map several synsets of the Indian language WordNets to the available images. A total of **8,183** Hindi synsets for directly linked nouns were mapped to their corresponding images. We perform manual evaluation of the data using the tool mentioned above and have evaluated approx. 3,000 synsets for each of the languages. We continue with the manual evaluation for mapping as of now.

Table 1 describes the number of synsets of the WordNets of the following languages for which images have been found, the number of evaluated images out of these, the correctly mapped images, and the precision score for each language.

The top three images are shown to a trained linguist who decides the winner image and also calculates the precision (P@3) of results for that synset. Over a set of 8183 images, we obtain a precision (P@3) of 0.30.

Algorithm 1 Image scoring algorithm

```
1: procedure IMAGE–SCORING
2:     score:= 0
3:     weight(ImageTags) := w
4:     cost(ImageTags) := c
5:     for each token i ∈ ImageTags do
6:         for each token j ∈ Synset do
7:             if i = j then
8:                 score:= score + w
9:             else
10:                score:= score - c
11:            end if
12:        end for
13:    end for
14: end procedure
```

6 Evaluation Tool

We create a PHP[7] based interface, and provide it to lexicographers and linguists for evaluation of the images obtained. The tool uses MySQL[8] database at the back-end to store both Hindi and English WordNet databases, and uses synset ID as a pivot to display the images obtained. The tool provides with a Hindi synset words, its concept, and the English words to help the lexicographer identify its proper sense. The lexicographer chooses a winner image out of the top three, or none of

[6] https://openclipart.org/search/json/

[7] http://php.net/

[8] https://www.mysql.com/

Languages	Images Obtained	Evaluated Images	Accurate Images	Precision
Hindi	8183	3851	1154	**0.3**
Assamese	5198	2860	771	0.27
Bengali	7823	3851	1154	**0.3**
Bodo	5138	2835	765	0.27
Gujarati	7736	3787	1134	**0.3**
Kannada	5695	2883	870	**0.3**
Kashmiri	6705	3470	1043	**0.3**
Konkani	7548	3686	1110	**0.3**
Malayalam	6504	3427	954	0.28
Manipuri	5299	2907	780	0.27
Marathi	6863	3452	1031	**0.3**
Nepali	3959	2163	584	0.27
Sanskrit	7812	3851	1154	**0.3**
Tamil	7272	3611	1083	**0.3**
Telugu	5728	2980	834	0.28
Punjabi	5889	3186	896	0.28
Urdu	5096	2683	684	0.25
Oriya	7412	3660	1034	0.28

Table 1: No. of synsets linked to images

Figure 2: Screen-shot of the Evaluation Tool

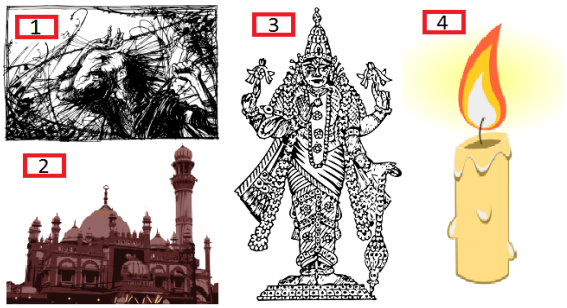

Figure 3: Accuractely acquired images

these, in case of no relevant image. They were also requested to tick the relevant images. A screen-shot for our interface is shown in Figure 2.

7 Qualitative Analysis

In this section, we explain the work done to evaluate the resultant images and the analysis of the results.

7.1 No images found

From the 11,573 sysets that were chosen to be tagged with images, We were unable to retrieve images for 3390 synsets from OpenClipArt, due to unavailability in the source. Our analysis shows that most of the synsets for which a suitable image could not be retrieved fell into two major categories:

Abstract nouns: Several of the synsets for which no images could be retrieved fell into the category of abstract nouns. For example the synset "गुलछर्रा" (*"gUlchharra"*) - 4939 which translates to "profligacy, extravagance" returned no images.

Complex synsets: Apart from abstract nouns, several complex synsets returned no results. For example, the synset "शारीरिक तरल पदार्थ" (*"shAririk TaRal PadArth"*) - 1644 which translates to "Liquid body substance" was unable to fetch any results.

We believe that synsets falling into the first category, *i.e* Abstract nouns, were too vague for an image to do justice to the concept. However, synsets falling into the second category display the limitedness of the OpenClipArt database and further the need of looking in more than one image source.

7.2 Images found

Amongst the synsets for which some images were retrieved, a link was noticed between the class of the noun and how well the image was able to explain the synset.

7.2.1 Common Nouns

Our methodology performs well in this case, and most of the images obtained were able to correctly and almost completely explain the concept. For example, the synset "मोमबत्ती" (*"momBatti"*) - 9866 meaning "candle" and synset "मस्जिद" (*"maszid"*) - 2900 meaning "mosque" retrieved very accurate results as shown in figures 3.4 and 3.2, respectively.

7.3 Proper Nouns

Our retrieval performs well for proper nouns. We were able to obtain pictures for most of the synsets which represent a country. The country flag and map was retrieved for each country name. Several Indian monuments obtained good images along with several Hindu deities. The illustration for synset "विष्णु" ("viShnU") - 2185 translating to a named entity "Vishnu" is shown in figure 3.3.

7.4 Abstract Nouns

Several images were unable to illustrate their corresponding abstract nouns. A few cases of good images were obtained such as synset "हड़कंप" (*"HaDKamp"*) - 3366 meaning "panic" was illustrated by the image 3.1.

8 Conclusion and Future Work

We successfully identified images for synsets of Indian languages and described our work on enriching IndoWordNet. Many synsets could not be linked due to the lack of appropriate image availability on OCAL. We also created a tool for manual evaluation of the data, or any other such work in the future. We evaluated the images obtained, and reported the highest precision score as 0.30. As a future work, we aim to try to retrieve these images using other open source image databases, and utilizing gloss and examples for finding overlaps. Also, The concept of Content Based Image Retrieval (CBIR) appears to be a viable option of several Indian language synsets which cannot be directly linked to a single corresponding English synset. Using CBIR, we can harness resources of several untagged image databases, and thus further enich IndoWordNet as a resource.

9 Acknowledgment

We gratefully acknowledge the support of the Department of Electronics and Information Technology, Ministry of Communications and IT, Government of India. We also acknowledge the annotation work done in this task by Rajita Shukla, Jaya Saraswati, Meghna Singh, Laxmi Kashyap, Ankit, and Amisha. Also, not to be missed, is the entire computational linguistics group at CFILT, IIT Bombay, which has provided its valuable input and critique, helping us refine our task.

References

Pushpak Bhattacharyya. 2010. Indowordnet. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, ALR7, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Prabhakar Pande Dipak Narayan, Debasri Chakrabarti and Pushpak Bhattacharyya. 2002. An experience in building the indowordnet - a wordnet for hindi. In *Proceedings of the First International Conference on Global WordNet (GWC'02)*, Mysore, India, January.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Jonathan Phillips. 2005. Introduction to the openclip art library.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.